DOCUMENT RESUME

ED 453 282 TM 032 819

AUTHOR Leung, Chi-Keung; Chang, Hua-Hua; Hau, Kit-Tai
TITLE Integrating Stratification and Information Approaches for Multiple Constrained CAT.
PUB DATE 2001-04-00
NOTE 21p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Seattle, WA, April 11-13, 2001).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE EDRS Price MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Item Banks; *Selection; Test Construction; *Test Items
IDENTIFIERS *Stratification

ABSTRACT

It is widely believed that item selection methods using the maximum information approach (MI) can maintain high efficiency in trait estimation by repeatedly choosing high discriminating (alpha) items. However, the consequence is that they lead to extremely skewed item exposure distribution in which items with high alpha values becoming overly exposed while many low alpha ones may never be selected. With some demonstrated empirical success, the alpha-stratified design (ASTR) and its extension, the alpha-stratified with b-blocking method (BASTR), were proposed to tackle these problems simultaneously. While the latter method takes into consideration the correlation between the alpha- and beta-parameters, both stratified designs advocate the more frequent use of low alpha items in earlier stages of testing. The main objective of this study was to investigate the performance of integrating the traditional information based selection and the new philosophy of using low alpha items first in a multiple constrained setting. The performances of MI, BASTR, and their integration, MIBASTR, were empirically compared through simulation. Results indicate that BASTR was the best in utilizing the entire pool evenly, and thus in tackling item security problems. On the other hand, MI and MIBASTR offered high and comparable measurement efficiency. The latter outperformed the former in item exposure and pool utilization. (Contains 2 figures, 4 tables, and 26 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 453 282

# Integrating Stratification and Information Approaches for Multiple Constrained CAT

Chi-Keung LEUNG
*The Hong Kong Institute of Education*

Hua-Hua CHANG
*National Board of Medical Examiners*

Kit-Tai HAU
*The Chinese University of Hong Kong*

*Note. Paper presented at the National Council of Measurement in Education Annual Meeting, Seattle, April 11-13, 2001.* *Address all correspondence to Chi-Keung Leung, Dept. of Math, Hong Kong Institute of Education, Taipo, N.T., Hong Kong. Email: ckleung@ied.edu.hk.*

TM032819

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY HAS BEEN GRANTED BY

C.-K. Leung

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

**Abstract**

It is widely believed that item selection methods using the maximum information approach (MI) can maintain high efficiency in trait estimation by repeatedly choosing high discriminating (*a*) items. However, the consequence is that they lead to extremely skewed item exposure distribution in which items with high *a* values become overly exposed while a lot of low *a* ones may never be selected. With some demonstrated empirical success, the *a*-stratified design (ASTR) and its extension, the *a*-stratified with *b*-blocking method (BASTR), were proposed to tackle these problems simultaneously. While the latter method takes into consideration the correlation between the *a*- and *b*-parameters, both stratified designs advocate the more frequent use of low *a* items in earlier stages of testing. The main objective of this study was to investigate the performance of integrating the traditional information based selection and the new philosophy of using low *a* items first in a multiple constrained setting. Specifically, the performances of MI, BASTR, and their integration, MIBASTR, were empirically compared. Results indicate that BASTR was the best in evenly utilizing the entire pool and hence tackling item security problems. On the other hand, MI and MIBASTR offered high and comparable measurement efficiency. But the latter outperformed the former in item exposure control and pool utilization.

# Integrating Stratification and Information Approaches for Multiple Constrained CAT

In the past decade, many large-scaled testing programs have been partly or completely converted into the computerized adaptive test (CAT) format that was first developed under the item response theory framework (IRT; Lord, 1970). In CAT, examinees are presented with tailor-made tests. One item is selected at a time on the basis of the examinee's currently estimated ability (Lord, 1980; Weiss, 1982). One of the main advantages of CAT over paper-and-pencil tests is that it enables more efficient and precise trait estimation (Owen, 1975; Wainer, 1990; Weiss, 1982). Other benefits include a larger flexibility in test scheduling and the incorporation of alternate item forms (Straetmans & Eggen, 1998). A key issue in CAT is how to adaptively select the best test items from the item pool. The traditional item selection algorithms rely on local item information. That means that an item is selected if it has the maximum Fisher information at the currently estimated ability based on the responses to previously administered items. It has been noted that this information criterion would cause skewed item exposure (Davey & Parshall, 1995; McBride & Martin, 1983; Stocking & Lewis, 1995; Sympson & Hetter, 1985; Thomasson, 1995; van der Linden, 1998). In particular, items with large value of discrimination parameter may be overly exposed while some others are never used. Over-exposure would eventually threaten in item security while under-utilization would reduce the cost effectiveness of developing and maintaining those inactive items in the pool. It is understandable, therefore, the control of item exposure and the enhancement of pool efficiency are important issues in CAT designs (Mills & Stocking, 1996; Stocking & Swanson, 1998; Way, 1998).

## The Stratification Designs

Methods that attempt to simultaneously control maximum item exposure rate to improve item security and uplift the exposure of under-utilized items to enhance item pool efficiency have been proposed by Chang and Ying (1999) and Stocking and Swanson (1998), among others. In contrast to the traditional approach of using the most informative

items at every stage of item selection, Chang and Ying (1999) proposed the multi-stage a-stratified design (ASTR) that partitions items into several strata in an ascending order of the item discrimination parameter. Each test then consists of matching numbers of stages and strata, with items of the first stage being selected from the first stratum, and so on. One major rationale for such a design is that at early stages, the gain in information by using the most informative item may not be realized because the ability estimation is still relatively inaccurate. Thus items with high discrimination values should be saved for later stages to accurately locate the ability. The ASTR has been empirically shown to be effective in both reducing item-overlap rate and the enhancing pool utilization with certain pool types. The findings of Hau and Chang (in press) also support that the ASTR offers comparable efficiency as MI but has the potential advantages of achieving a more balanced item usage and stable resultant pool structure in testing with continuous item replenishments.

For operational item pools, it is common to find substantial correlation between the *a*- and *b*- parameters of the items. Thus, at the later stages of testing with the high *a* strata, there may not be sufficient low *b* items. Consequently, the ASTR would result in an overexposure of these low *b* items. To tackle such problem, Chang, Qian and Ying (in press) modified the ASTR strategy and developed the *a*-stratified with *b*-blocking method (BASTR). In BASTR, item pool is first divided into many small levels based on *b* parameters. Within each level, items are further grouped in ascending order of *a* values. The first group of items with the smallest *a* values from each level are then merged to form the first stratum, and the second group of items with the second smallest *a* values merged into the second stratum and so on, until eventually the last stratum is formed from items with the largest *a* values from each level. As a result, the *b* distribution remains steady but the average *a* value increases across strata.

Addressing Multiple Constraints

Very often, CAT design has to take into consideration additional constraints such as content balancing, item type and statistical specifications. Various models have been

proposed to solve such complex and sometimes conflicting requirements in test assembly (Armstrong, Jones & Kunce, 1998; Luecht, 1998; Stocking & Swanson, 1993, 1998; van der Linden & Reese, 1998). One strategy, as advocated by Stocking and Swanson (1993) in the Weighted Deviation Model (WDM) is to relax test constraints as desired properties. It can deal with many kinds of constraints and provide acceptable solution to those systems that may not have an optimal combination satisfying all criteria. Leung, Chang, and Hau (2000) have shown that the two integrated item selection methods, BASTR-WDM and ASTR-WDM, to a certain extent, can tackle the problem of multiple constraints. However, they have also pointed out that more thorough examinations on pool partition and item selection are needed to guarantee face validity and measurement efficiency. One major problem as in BASTR-WDM and ASTR-WDM designs is that it may not be easy to find suitable items in the later strata to satisfy the unfulfilled constraints towards the end of the adaptive testing.

In this study, we examine the possible strengths of the stratification design and the information approach for multiple constrained CAT. Specifically, an item pool is first partitioned into two strata based on the pool characteristics using the general concept of BASTR. As a result, both strata cover a wide range of *b*s with the first stratum having a smaller *a* average. The test is then divided into two stages. As the trait estimation is generally inaccurate in the early stage of a test, the items of the first stage are selected from the first stratum by matching the *b*-parameter and the current ability estimate. Whereas in the final stage, items are selected from the second stratum based on the maximum information approach to accurately locate the true theta. It is suggested that the first stage of a test is relatively shorter than the final stage. The main reasons of doing this are: to uplift the utilization of low *a* items, to increase the availability of suitable items for unfulfilled constraints during the final stage, and to increase the precision in trait estimation. In the study, the performance of the new integrated method, MIBASTR, was examined and compared with the BASTR and MI methods.

## Simulation Studies

A series of simulation studies were conducted to investigate the performance of MI, BASTR, and MIBASTR, when multiple constraints were imposed. There were 19 non-statistical constraints arising from two intrinsic features: content area and cognitive level. The sample consisted of 5,000 examinees with abilities (referred as true ability $\theta$ hereafter) randomly generated from N(0,1). Each examinee received three simulated tests administered respectively by the three item selection methods incorporated with WDM: MI, BASTR, and MIBASTR.

### Item Pool

The item pool imitated the psychometric characteristics of an operational pool containing 700 upper primary mathematics items with four content areas crossed with three cognitive levels. Table 1 shows the distribution of the items. The mean estimated discrimination ($a$) value for the items was 1.02, with SD = .33, and a range of .29 to 2.63. The mean estimated difficulty ($b$) was .16, with SD = 1.06, and a range of –3.44 to 3.40; thus there was a close match between the $b$ distribution and ability distribution. The mean estimated pseudo-guessing parameter ($c$) was .17, with SD = .008, and a range of .03 to .50. The correlation of $a$- and $b$-parameters was .45.

==========

Insert Table 1 about here

==========

### Test Constraints

The test length was fixed at 42 items. Table 2 presents the desired ranges and relative weights for various combinations of content area and cognitive level.

==========

Insert Table 2 about here

==========

### Item Selection Methods

The performances of the following three item selection methods were compared when WDM was incorporated to deal with multiple constraints. In the WDM, test constraints are treated as desired properties. Thus, when there is no feasible solution to the system of the constraints, the objective function could find an item yielding the minimal deviation from the constraints. The model used in the study is:

Minimize

$$\sum w_j d_{L_j} + \sum w_j d_{u_j} + w_\theta d_\theta \qquad \text{(sum of weighted deviations)} \qquad (1)$$

Subject to

$$\sum_{i=1}^{N} a_{ij} x_i + d_{L_j} - e_{L_j} = L_j \quad , j = 1,\ldots,19 \qquad \text{(for lower bounds of constraints)} \quad (2)$$

$$\sum_{i=1}^{N} a_{ij} x_i + d_{U_j} - e_{U_j} = U_j \quad , j = 1,\ldots,19 \qquad \text{(for upper bounds of constraints)} \quad (3)$$

$$\sum_{i=1}^{N} I_i(\theta) x_i + d_\theta - e_\theta = \infty \qquad \text{(for max. information criterion)} \quad (4)$$

$$d_{L_j}, d_{U_j}, e_{L_j}, e_{U_j} \geq 0 \quad , j = 1,\ldots,19 \qquad (5)$$

and

$$x_i \in \{0, 1\}, i = 1,\ldots, N, \qquad \text{(Stocking \& Swanson, 1993, p. 280-281)} \qquad (6)$$

where $a_{ij}$ equals to 1 if item $i$ has property $j$ and 0 otherwise; $w_j$ is the weight assigned to constraint $j$; $w_\theta$ is the weight assigned to the information constraint; $L_j$ and $U_j$ are the lower bound and upper bound of constraint $j$; $d_{L_j}$ and $d_{U_j}$ represent deficit from the lower bound and surplus from upper bound respectively; $e_{L_j}$ and $e_{U_j}$ represent excess from lower bound and deficit from upper bound respectively; $x_i$ equals to 1 if $i$th item is included in the test, or 0 otherwise.

The heuristic for item selection using WDM is also adopted:

1. For every item not already in the test, compute the deviation for each of the constraints if the item is added to the test;

2. Sum the weighted deviations across all constraints;
3. Select the item with the smallest weighted sum of deviations (Stocking & Swanson, 1993, p. 281).

*Method 1 (MI)*:

1. The content area and cognitive level of each item are indexed.
2. The optimal function of Equation 1 is used.
3. Test constraints and maximum information are treated as desired properties as stated in Equations 2 to 4.
4. The bounds and weights for the 19 non-statistical constraints are set as stated in Table 2. The lower bound for information is set at 100.0 which is practically unreachable in the study. The weight for information is assigned the value of 15.0 as has been used by Stocking and Swanson (1993) and found to have yielded very small mean squared error.
5. Items are selected by following the three steps of the heuristic mentioned earlier.

*Method 2 (BASTR)*:

1. The content area and cognitive level of each item are indexed.
2. Test constraints are mathematically formulated as Equations 2 and 3.
3. The item pool is partitioned into 4 strata as follows. First, the pool is divided into 20 different levels based on the *b*-parameters. Then the items within each level are sorted in ascending order of *a*-parameters and then divided into 4 groups. All the first groups of the 20 levels are merged to form the first stratum;...; and all the last groups merged to form the last stratum. As a result, each stratum then has similar distribution for *b*-parameters but the average value of *a*-parameters increase across the strata. Table 3 shows the parameter distributions across strata after stratification.
4. Each test is divided into 4 stages. In the first stage, 10 items are administered from the first stratum. In the second stage, the next 10 items are administered from the second stratum and then 11 items from each of the last two strata.
5. At each point of item selection, a hundred unadministered items with difficulty

parameter closest to the currently estimated theta are chosen. For each of them, the absolute difference between the *b* value and the current ability estimate is computed. The weight for such absolute difference is set at 5.0. Then the total weighted deviations are summed up as if the item has been added to the test. The item with the smallest weighted sum of deviations is administered.

==========

Insert Table 3 about here

==========

*Method 3 (MIBASTR)*:

1. The constraints and information criterion are formulated as in Equations 2 to 4.
2. The item pool is partitioned into two levels. The first level corresponds to the first stratum while the second level is formed by combining the last three strata of BASTR.
3. Accordingly, each test is divided into 2 stages with the first 10 items from the first stage and the remaining 32 items from the final stage.
4. The item selection algorithm in the first stage is the same as described in BASTR.
5. In the final stage, all unadministered items of the second stratum are considered. For each of them, the deviation for each of the desired properties, including information, is computed as if the item is added to the test. The weighted deviations across all constraints are summed and then the item with smallest weighted sum of deviations will be administered.

## Measures of Performance

Reliability

Irrespective of the item selection design, CAT should always provide reasonable reliability. Otherwise, test results cannot be used for inference or decision. Therefore, reliability was an evaluation criterion for the performance of the three selection methods. In this study, 5,000 true abilities were generated and their respective estimates were obtained according to the selection methods employed. Thus reliability here was defined as the

correlation between the estimated scores and the true scores (Lord, 1980, p. 52). The higher the reliability, the better the item selection method would be.

Bias

Accuracy is another important criterion, which in this study was measured by the estimated bias and mean squared error. Let $\theta_i$, $i = 1,\ldots, 5000$ be the true abilities of the 5000 examinees and $\hat{\theta}_i$ be the respective estimators from the CAT. Then the estimated bias was computed as

$$\text{Bias} = \frac{1}{5000}\sum_{i=1}^{5000}(\hat{\theta}_i - \theta_i) \qquad (7)$$

The smaller the bias, the better the item selection method would be.

Mean Squared Error

Using the same notations of true ability and the estimator, the estimated mean squared error was computed as

$$\text{MSE} = \frac{1}{5000}\sum_{i=1}^{5000}(\hat{\theta}_i - \theta_i)^2 \qquad (8)$$

The smaller the MSE, the better the item selection method would be.

Number of Over-exposed Items

The exposure rate of an item is defined as the ratio of the number of times the item is administered to examinees over the total number of examinees taking the test. If an item has a high exposure rate, then it has a greater risk of being known to prospective examinees, which in turn would cause test security and validity problems. Since one of the main concerns of this paper was to compare the three item selection methods in item exposure control, the number of overly exposed items was certainly a key evaluation criterion. Here an item was considered overly exposed if its exposure rate was greater than .2, a commonly used cut-off value. The smaller the number of over-exposed items, the better the item

selection method would be.

Number of Under-utilized Items

Items with very low exposure rate are rarely used. If there are too many items with low exposure rates, then the item pool is not well utilized, which challenges directly the cost effectiveness of the item pool and the appropriateness of the item selection method. In this study, an item was considered as under-utilized if its exposure rate was below .02. The smaller the number of under-utilized items, the better the item selection method would be.

Scaled Chi-squared Statistic

Chang and Ying (1999) proposed that a uniform exposure rate distribution should be the most desirable in order to have a maximum item pool utilization. If the pool size is $N$ and test length is $L$, then the optimum uniform exposure rate is $L/N$. They introduced a scaled chi-square to measure the overall item pool usage efficiency:

$$\chi^2 = \sum_{j=1}^{N} \frac{(er_j - L/N)^2}{L/N} \qquad (9)$$

where $er_j$ represents the observed exposure rate for the $j$th item.

Equation 9 reflects the discrepancy between the observed and the ideal exposure rates. The smaller the $\chi^2$, the better the pool utilization and hence the item selection method would be.

Item Overlap Rate

The item overlap rate (sometimes called the test overlap rate) is another important summary index in measuring item exposure control (Mills & Stocking, 1996; Way, 1998). Item overlap rate is indicated by the proportion of items shared by pairs of examinees, averaged across all possible pair-wise combinations. Way (1998) argued that such an index, not being calculated on an individual item basis, provides a global picture of how

often sets of items are administered. The higher the item overlap rate, the greater the damage to test validity due to information sharing among examinees who take the test at different occasions. He also stressed that this index is critical in determining the size and composition of item pools that are needed for a particular CAT. If the test length is $L$ and there are $P$ examinees, the item overlap rate here was computed by: (i) first counting the number of common items for each of the $P(P\text{-}1)/2$ pairs of examinees, (ii) adding up all the counts in the $P(P\text{-}1)/2$ pairs, and (iii) dividing the total count by $LP(P\text{-}1)/2$. The smaller the overlap rate, the better the item selection method would be. Chang and Zhang (1999) and Chen, Ankenmann and Spray (1999) separately found that the lower bound for the expected item-overlap rate is $L/N$, the ratio of the test length to the pool size. This lower bound serves as a baseline for comparison among item selection methods in controlling item overlap.

Deviations from Desired Ranges

As the desired ranges are usually recommended by experts of the relevant fields, it is anticipated that the CATs by all selection methods should fall within these ranges. The smaller the deviations from the desired ranges of the non-statistical constraints, the higher the face validity and thus the better the item selection method would be.

## Results

The results of the study are summarized in Table 4. The three item selection methods appeared virtually unbiased as their estimated biases were all close to zero. They offered high and comparable reliabilities. BASTR outperformed the other two methods in maintaining test security and even pool utilization. It greatly reduced the risk of item leakage by substantially lowering item overlap rate and the number of over-exposed items. In addition, it made full use of the item pool as indicated by their smaller values in both the chi-square statistic and the number of under-utilized items.

==========

Insert Table 4 about here

==========

However, BASTR alone without the maximum information strategy seemed inadequate in its measurement efficiency. Its MSE was larger than that of the other two methods. Both MI and MIBASTR were more efficient and they yielded comparable MSE. Regarding item security, MIBASTR outperformed MI in having less overly exposed items and a smaller overlap rate. Besides, the observed maximum exposure rate was smaller. It was also superior in item pool utilization as indicated by its smaller chi-square value and less under-utilized items.

All the three methods conformed well to the multiple constraints with at most one observed deviation among all tests. BASTR had the smallest number of testing with deviation whilst MI had the most number of such occurrence.

Figure 1 shows the item exposure distributions for the three methods. The curve for BASTR indicates that the method utilized items mostly with a maximum exposure slightly over .3. In contrast, MI left about 60% of the pool items untouched, thus raising the issue of the cost effectiveness of developing and maintaining the pool. MIBASTR had similar problems but with only about 45% of items never used. Besides, the item security problem in MIBASTR was less serious than that in MI with the former having less over-exposed items.

==========

Insert Figure 1 about here

==========

Figure 2 presents the individual exposure rates of items sorted in an ascending order of *a*-parameter within each stratum. It is clear that MI repeatedly selected highly discriminating items that were near the end of each stratum, while under-utilizing a lot of low *a* items around the front of stratum. BASTR yielded a much even exposure distribution and utilized the pool more effectively. Figure 2(c) reflects that MIBASTR

tended to equalize the item exposures in the first stratum of which items were selected by matching *b* with the estimated theta. But in the second stage that involved items from the last three strata, MIBASTR yielded similar exposure pattern as MI did. It appears that imposing the stratified design could partly solve the over-exposure and poor pool utilization problems associated with MI.

==========

Insert Figure 2 about here

==========

## Discussion

There are two common threats associated with the information-based item selection methods, namely, poor item pool utilization and test security problems. Both ASTR and BASTR were developed in an attempt to remedy these problems; the latter has been shown to be more appropriate for item pools in which *a*- and *b*- parameters are strongly correlated. These two stratified methods can work effectively with WDM to meet multiple constrained requirements (Leung, Chang & Hau, 2000). However, they on their own are not as efficient as MI is. The main objective of this study was to investigate how the integration of MI and BASTR would perform in a multiple constrained setting.

Results indicate that the new philosophy advocated in the stratification designs, using low *a* items first, can be combined with the traditional wisdom of information-based criterion. It has been shown that both MI and MIBASTR offered comparable measurement efficiency and accuracy. But the latter outperformed the former in dealing with the practical issues of pool utilization and item security. MIBASTR demonstrated two important strengths of the stratification. Firstly, it enhances pool utilization by reducing the number of under-utilized items and the value of scaled chi-square statistic. Secondly, it improves item security by lowering the maximum exposure rate, the number of over-exposed items, and the item overlap rate. The integrated method also carries the efficiency of information approach by yielding similar MSE and reliability as MI does.

The current findings have two implications for CAT design and pool management. First, there is no conflict between the new philosophy of stratification and the traditional wisdom of using information criterion. The two concepts can be combined to form an effective method for addressing practical issues such as pool utilization, item security and measurement efficiency. Second, using low *a* items and *b*-matching criterion at the beginning of a test would not affect the measurement efficiency due to a considerable discrepancy between the true theta and its initial stage estimate. In fact, this strategy tends to equalize the exposure rates of those items in the first stratum. Thus, if there are some inactive items in a testing program, their usage can be uplifted by streaming them into the first stratum. This would improve the cost effectiveness of the development and maintenance of these items, without sacrificing the measurement efficiency.

The integration of MI and BASTR has been shown to be more promising than either one alone. It should be noted that though the over-exposure problem of MI has been lessened in MIBASTR, the number of over-exposed items is still high. Furthermore, future research may investigate how the integration of stratification design and information approach would perform when some additional measures of exposure control are imposed. Other factors that may affect the performance include the size and the number of strata, the number of constraints and the corresponding ratios, the matching of *b* distribution and the ability population, and the test termination rule.

## References

Armstrong, R.D., Jones, D.H., & Kunce, C.S. (1998). IRT test assembly using network-flow programming. *Applied Psychological Measurement, 22*, 237-247.

Chang, H.H., Qian, J., & Ying, Z. (in press). *a*-Stratified Multisage CAT with b-Blocking. *Applied Psychological Measurement*.

Chang, H.H., & Ying, Z. (1996). A Global Information Approach to Computerized Adaptive Testing. *Applied Psychological Measurement, 20*, 213-229.

Chang, H.H., & Ying, Z. (1999). A-stratified Multistage Computerized Adaptive Testing. *Applied Psychological Measurement, 20*, 213-229.

Chang, H.H., & Zhang, J. (1999, June). *Hypergeometric family and test overlap rates in computerized adaptive testing*. Paper presented at the annual meeting of the Psychometric Society, Lawrence, KS.

Chen, S., Ankenmann, R.D., & Spray, J.A. (1999, April). *Exploring the relationship between item exposure rate and test overlap rate in computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Davey, T., & Parshall, C.G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, USA.

Hau, K.T., & Chang, H.H. (in press). Item Selection in Computerized Adaptive Testing: Should More Discriminating Items be Used First?. *Journal of Educational Measurement*.

Leung, C.K., Chang, H.H., & Hau, K.T. (2000). *Solving Complex Constraints in a-Stratified Computerized Adaptive Testing Designs*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, USA.

Lord, M.F. (1970). Some test theory for tailored testing. In W.H. Holzman (Ed.), *Computer Assisted Instruction, Testing, and Guidance*. New York: Harper and Row.

Lord, M.F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Luecht, R.M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, *22*, 224-236.

McBride, J.R., & Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.

Mills, C.N., & Stocking, M.L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, *9*, 287-304.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351-356.

Stocking, M.L., & Lewis, C. (1995). *A new method of controlling item exposure in*

*Computerized Adaptive Testing*. Research Report 95-25. Princeton, NJ: Educational Testing Service.

Stocking, M.L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277-292.

Stocking, M.L., & Swanson, L. (1998). Optimal design of item banks for computerized adaptive tests. *Applied Psychological Measurement, 22*, 271-279.

Straetmans, G.J., & Eggen, T.J. (1998). Computerized adaptive testing: what it is and how it works. *Educational Technology, 38*, 45-52.

Sympson, J.B., & Hetter, R.D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th Annual Meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thomason, G.L. (1995, June). *New item exposure control algorithms for computerized adaptive testing*. Paper presented at the Annual Meeting of Psychometric Society, Minneapolis, MN.

van der Linden, W.J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201-216.

van der Linden, W.J., & Reese, L.M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259-270.

Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum.

Way, W.D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17*, 17-27.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

Table 1: Item Distribution across Content Areas and Cognitive Levels

| | Level 1 | Level 2 | Level 3 |
|---|---|---|---|
| Content A | 52 | 142 | 40 |
| Content B | 33 | 104 | 29 |
| Content C | 27 | 93 | 30 |
| Content D | 28 | 81 | 41 |

Table 2: Ranges and Weights for 19 Non-statistical Constraints

| | Lower Bound | Upper Bound | Weight |
|---|---|---|---|
| Content A | 10 | 17 | 40 |
| Content B | 8 | 12 | 40 |
| Content C | 7 | 11 | 40 |
| Content D | 7 | 11 | 40 |
| Level 1 | 7 | 11 | 40 |
| Level 2 | 19 | 31 | 40 |
| Level 3 | 7 | 11 | 40 |
| Content A, Level 1 | 2 | 5 | 7 |
| Content A, Level 2 | 6 | 11 | 7 |
| Content A, Level 2 | 2 | 4 | 7 |
| Content B, Level 1 | 1 | 3 | 7 |
| Content B, Level 2 | 4 | 8 | 7 |
| Content B, Level 3 | 0 | 3 | 7 |
| Content C, Level 1 | 0 | 3 | 7 |
| Content C, Level 2 | 3 | 7 | 7 |
| Content C, Level 3 | 0 | 3 | 7 |
| Content D, Level 1 | 0 | 3 | 7 |
| Content D, Level 2 | 4 | 8 | 7 |
| Content D, Level 3 | 2 | 4 | 7 |

Table 3: Parameter Distributions across Strata of BASTR

| Stratum | | *Discrimination (a)* | *Difficulty (b)* | *Guessing (c)* |
|---|---|---|---|---|
| 1 | Mean | .70 | .16 | .18 |
| | S.D. | .17 | 1.11 | .008 |
| 2 | Mean | .92 | .16 | .17 |
| | S.D. | .17 | 1.04 | .008 |
| 3 | Mean | 1.10 | .17 | .18 |
| | S.D. | .20 | 1.07 | .008 |
| 4 | Mean | 1.38 | .17 | .17 |
| | S.D. | .31 | 1.04 | .008 |

Table 4: Summary Statistics for Three Selection Methods

| | MI | BASTR | MIBASTR |
|---|---|---|---|
| Bias | -.0051 | -.0035 | -.0036 |
| MSE | .034 | .059 | .036 |
| Reliability | .98 | .97 | .98 |
| Scaled $\chi^2$ | 195.8 | 16.6 | 122.0 |
| N(exp<.02) | 472 | 60 | 362 |
| N(exp>.2) | 86 | 7 | 63 |
| Min exp | .000 | .000 | .000 |
| Max exp | .630 | .317 | .540 |
| Overlap Rate | .274 | .079 | .225 |
| *Deviations from Non-statistical Constraints* | | | |
| N(1 dev) | 201 | 127 | 169 |
| N(2 or above) | 0 | 0 | 0 |

Figure 1: Item Exposure Distributions for Three Selection Methods

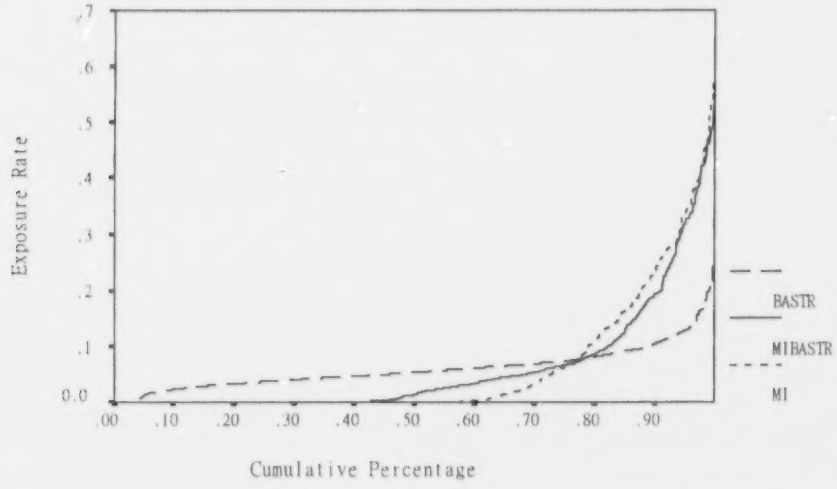


Figure 2: Individual Item Exposure Rates for Three Selection Methods

(a) MI (b) BASTR (c) MIBASTR